Quantitative Evaluation of a Pulmonary Contour Segmentation Algorithm in X-ray Computed Tomography Images[1]

Beatriz Sousa Santos, PhD, Carlos Ferreira, PhD, José Silvestre Silva, MSc, Augusto Silva, PhD, Luísa Teixeira, MD

Rationale and Objectives. Pulmonary contour extraction from thoracic x-ray computed tomography images is a mandatory preprocessing step in many automated or semiautomated analysis tasks. This study was conducted to quantitatively assess the performance of a method for pulmonary contour extraction and region identification.

Materials and Methods. The automatically extracted contours were statistically compared with manually drawn pulmonary contours detected by six radiologists on a set of 30 images. Exploratory data analysis, nonparametric statistical tests, and multivariate analysis were used, on the data obtained using several figures of merit, to perform a study of the interobserver variability among the six radiologists and the contour extraction method. The intraobserver variability of two human observers was also studied.

Results. In addition to a strong consistency among all of the quality indexes used, a wider interobserver variability was found among the radiologists than the variability of the contour extraction method when compared with each radiologist. The extraction method exhibits a similar behavior (as a pulmonary contour detector), to the six radiologists, for the used image set.

Conclusion. As an overall result of the application of this evaluation methodology, the consistency and accuracy of the contour extraction method was confirmed to be adequate for most of the quantitative requirements of radiologists. This evaluation methodology could be applied to other scenarios.

Key Words. Quantitative evaluation; computed tomography (CT); pulmonary segmentation; interobserver and intraobserver variability.

We have reached a point at which computed tomography (CT) images can be reconstructed faster than they can be read. This fact encourages software developers to design programs that will aid radiologists in the reading of CT images and in diagnosing conditions on the basis of CT findings (1). Segmentation often occurs as a preprocessing step of more global image analysis tasks, as is the case of computer-aided analysis of pulmonary x-ray tomograms (2), where many analytic procedures start by correctly identifying the pulmonary regions (3–6). Most algorithms for the segmentation of pulmonary regions are based on intensity discrimination within the Hounsfield scale (7–9); however this task may become very complex because of

Acad Radiol 2004; 11:868–878

[1] From the Departamento de Electrónica e Telecomunicações (B.S.S., J.S.S., A.S.), the Instituto de Engenharia Electrónica e Telemática de Aveiro (B.S.S., A.S.), and the Departamento de Economia, Gestão e Engenharia Industrial (C.F.), Universidade de Aveiro, Portugal; the Centro de Investigação Operacional, Universidade de Lisboa, Portugal (C.F.); the Departamento de Física, Faculdade de Ciências e Tecnologia, Universidade de Coimbra, Portugal (J.S.S.); Serviço de Imagiologia, Hospitais da Universidade de Coimbra, Portugal (L.T.). Received January 14, 2004; revision requested March 30; revision received and accepted April 26. **Address correspondence to** B.S.S. DET Departamento de Electronica e Telecomunicaçoes, Universidade de Aveiro, 3810 Aveiro, Portugal.

doi:10.1016/j.acra.2004.05.004

the presence of spurious structures within the same scale range or the visual merging of the pulmonary regions themselves. In previous works (10,11) we presented algorithms designed to cope with these difficulties, which generate contours with a variable degree of similarity to those provided by radiologists.

A quantitative evaluation of the performance of these algorithms is crucial before their clinical use can be considered. Yet, the performance evaluation of segmentation algorithms in medical imaging is recognized as a difficult problem; actually, if one can find in the literature a significant number of contributions concerning the overall segmentation problem by itself, the same is not true when looking for quality and effectiveness assessments performed in some systematic way (12) and having a practical value (13).

This evaluation encounters the first great obstacle: the fact that the ground truth is unknown (13) (ie, it is not possible to identify the real contour corresponding to a given image). This problem is often circumvented using the contour resulting from manually tracing the object boundary by a knowledgeable human as a surrogate of that truth. However, not only will contours drawn by two radiologists be different (interobserver variability), but there will also not be agreement between contours drawn by the same radiologist at different occasions (intraobserver variability). These two types of variability have to be taken into account in the performance evaluation of segmentation algorithms; we will have to compare this performance with the performance of several radiologists in some statistically supported manner.

In an earlier work (11) we verified that a greater similarity existed between the contours produced by our algorithm and the contours drawn by two expert radiologists, than between the contours drawn by the same two radiologists. This meant that the interobserver variability between our algorithm and any of the two radiologists was less than the interobserver variability between the two radiologists. To investigate if this was specific for those two radiologists, or if it was more general, we have performed a study including six radiologists from different hospitals.

To further investigate this issue, we have considered the study of the intraobserver variability relevant; in this respect our algorithm has a clear advantage because its intraobserver variability is zero. Still, the comparison of the interobserver variability between our algorithm and each radiologist to hers/his intraobserver variability could provide interesting additional information on the performance of the algorithm.

While other authors have proposed pulmonary segmentation algorithms and have evaluated them (4,8), they have not compared their performance as contour detectors with as many radiologists, nor have they used such a statistically based method as we have used in this study.

MATERIALS AND METHODS

Quality Assessment Strategies

It is common to treat the physician ground truth as unquestionable, and assume it as a relatively error-free gold standard; however, there is some level of variability in the specification of the ground truth and it is important to have an estimate of this level. This type of variability is an important concern in determining the appropriate criteria for matching a detected contour to a ground truth contour (13).

Quantitative evaluation of the performance of segmentation algorithms in medical imaging has been recognized as an important problem. However, many of the evaluation studies that have been carried out did not use a large enough dataset, real images, convenient performance metrics, appropriate statistical methods, or a suitable ground truth. Thus, they cannot be considered correct or complete. Several methodologies have been proposed to perform this evaluation appropriately. The *Handbook of Medical Imaging* (13) presents a thorough overview of the field. Chalana and Kim (12) also present a concrete approach to segmentation performance assessment through contour comparison.

Quantitative Evaluation of the Performance

As mentioned previously, the ideal way of evaluating the performance of our segmentation algorithm would be to compare the contours detected on a valid test dataset with the "real contours" corresponding to each image. However, as we have seen, there are no such real contours. Several expert radiologists will detect different contours on the same image (see Fig 1b); also, each expert radiologist will detect on the same image, at different times, slightly different contours, unlike our algorithm, which always detects the same contours on the same image (its variability is 0 and it does not depend on any seed points introduced by a human observer, as other pulmonary segmentation algorithms (14–18)). This intraobserver variability can be used as a "variability quantum";

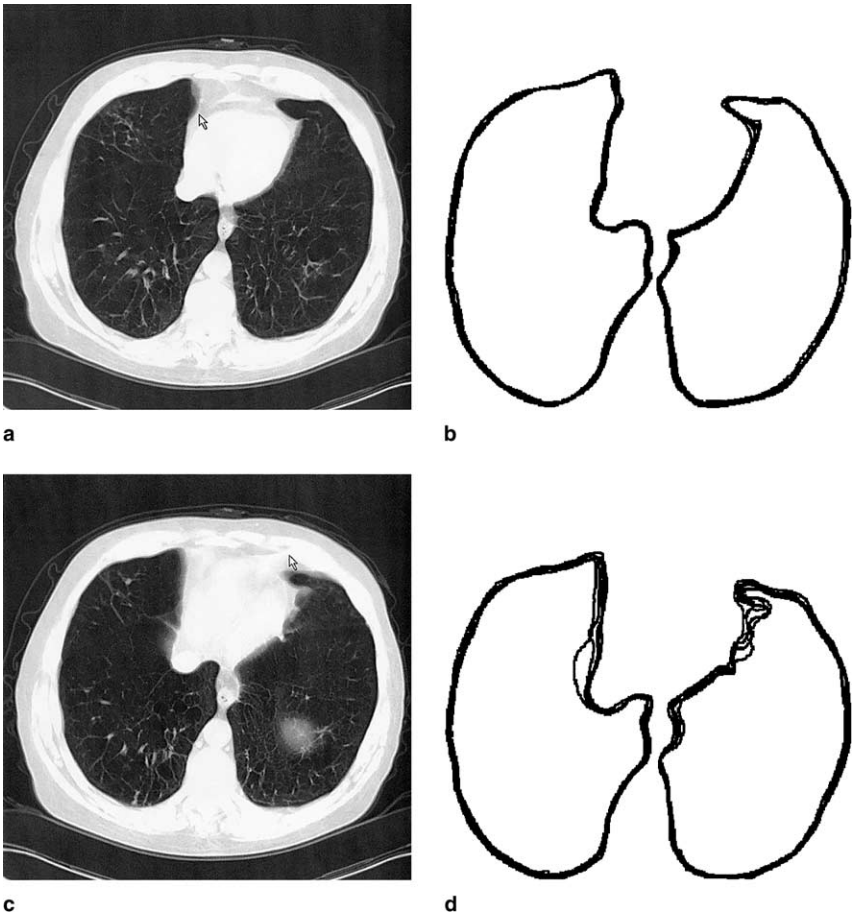

Figure 1. Images **(a, c)** and corresponding contours detected by six radiologists **(b, d)**.

it gives an idea of the level of variability that has to be expected and thus can be acceptable to exist between any two contour detectors (algorithm or human). Therefore, comparing the variability between our algorithm and each radiologist with the intraobserver variability of any expert could work as an "acceptability measure" of that variability.

As a consequence of the intraobserver and interobserver variability, the manually drawn contours can be considered as a collection of ground truths, all of them equally acceptable. To circumvent this problem we have performed two studies (using different methods) to compare the behavior of our algorithm, as a contour detector, with the behavior of a reasonable number of expert radiologists. These studies involve the assessment of the interobserver variability among a number of "contour detectors": several humans and one automated (our algorithm). The rationale for this study was that, if the interobserver variability between the algorithm and any of the radiologists is similar in magnitude to the interobserver variability between any two radiologists, then the difference between the algorithm and the radiologists, as contour detectors, could be considered not significant. This rational is similar to the one behind the study by Sivaramakrishna et al (19) to validate a segmentation algorithm of mammographic images, which also seems comparable to our case.

Interobserver Variability

In our first study concerning interobserver variability we directly compared the contours produced by all detectors (algorithm vs all radiologists and every radiologist vs all the others and algorithm). In a subsequent study we compared each detector with a reference contour (surrogate ground truth) obtained from the hand-drawn contours, as described by Ferreira et al (20).

To perform these studies we asked six experienced radiologists, from three different hospitals, to draw contours of the pulmonary regions on the chosen images. This number of radiologists seemed reasonable for such a

study, taking into consideration that they have been trained and work at three different hospitals. Moreover, it would be difficult to obtain the collaboration of more radiologists.

Intraobserver Variability

To assess intraobserver variability, we asked two radiologists to hand-draw the contours on the same set of images twice, without telling them that they had already drawn contours on those images. We chose the youngest radiologist and the head of the CT department who was responsible for thoracic radiology at the University Hospital, because these radiologists have a significant difference in years of experience. This choice was made in the hope of obtaining two significantly different values of intraobserver variability (which would probably not be the case if the two radiologists had approximately the same experience).

The time elapsed between the delineation of the two contours on the same image by the same radiologist was at least 1 month (which agrees with the proposal of Wagner et al (21)]) to minimize the effect of the recollection of having drawn the previous contours.

Test Dataset

The proper choice of the used dataset is very important; a poor selection of either the number of images or the method to select these images can jeopardize the validity of the evaluation procedure. We used 30 512 $\times$ 512 images ($N = 60$ contours) selected using a pseudorandom generator from a set of 253 images that had not been used to develop the algorithm. These images were all the images that could be used to support diagnosis corresponding to exams of eight patients collected at the Radiology Department of the University Hospital in Coimbra, independently of their pathologies. While the used dataset contained images corresponding to different pulmonary levels, which increased variability, using images from a greater number of patients would probably increase case variability. We used the power of a hypothesis test to calculate the sample size, N, of the test dataset, specifying the smallest difference that would be worthwhile to detect. This means, according to Altman (22), trying to make "clinical" importance and statistical significance agree. As a first approach, we hoped to be able to detect a difference of 1 standard deviation. We set the power (1-β) at 90% and chose a 1% significance level (α); using the nomogram for calculating sample size (22), this gives a total sample of $N = 60$.

Hand-outlined Contours

Our radiologists manually outlined all the contours on transparent sheets superimposed on quality printings of the test images working independently from each other and (as much as possible) in the same way and on the same conditions. The obtained contours were digitized and processed to identify the contours of left and right lungs. This identification is performed computing the image Radon transforms for 0° and 90°, estimating the center of each lung from the maximum values of these two transforms. Applying a morphologic filling starting from the center of one lung and a second filling starting from any point external to the lungs, we obtained an image containing the filled area of the other lung. The contour of the lung was then easily obtained. Erosion was applied to obtain a thinner version of each of the contours (20).

We have chosen this method as a compromise between feasibility to the radiologists and acceptable accuracy.

References Contour Obtained from the Hand-drawn Contours

In the last study of interobserver variability, we compared all contours to reference contours obtained from the six contours detected by the radiologists on each image, as described by Ferreira et al (20). For most images having diagnostic value, the contours detected by all the radiologists are only slightly different and thus using a kind of "average" contour seemed an acceptable surrogate to "ground truth" (Fig 1b); however, in particular regions of a few images of the data set, affected by partial volume effect or motion artifacts, the six radiologists detected contours that seem to correspond to the use of different segmentation criteria (Fig 1d); in this case an "average" contour does not make sense as "ground truth" and a different approach should be used. This needs further investigation; however, the impact on the results of this study is not expected to be significant because of the small number of images and reduced zones where this fact was observed in the used data set.

Comparing Contours

The comparison between any two contours was accomplished in two different ways: one based on the local distances between contours and the other exploring a similarity measure between the image masks (binary images containing the pulmonary areas defined by the contours).

The computation of distances between contours implies defining pairs of matching points on both contours. To

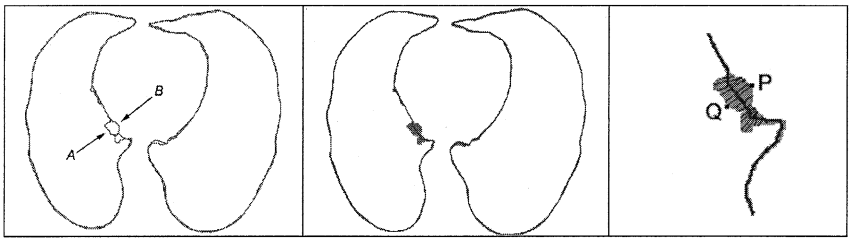

Figure 2. Definition of the auxiliary contour to obtain pairs of corresponding points on two contours A and B. *P* and Q are matching points on the contours under comparison.

find these pairs of points we used an auxiliary contour as shown in Figure 2. Differences between the contours were quantified using the Euclidean distances measured between corresponding points (11).

Figures of Merit

The values of the computed distances between the contours allow a localized and accurate quantification of their differences, easily assessable through simple visualization techniques. However, we consider it fundamental to use global quality figures of merit, which facilitate a comprehensive comparison. Thus, several figures of merit, based on the computed distances, were used as performance measures:

the Pratt figure of merit F_{Pratt} (23):

$$F_{\text{Pratt}} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{1+\alpha\times d_i^2} \tag{1}$$

the Mean Distance:

$$d_{\text{mean}} = \frac{1}{N}\sum_{i=1}^{N} d_i \tag{2}$$

the Maximum Distance:

$$d_{\text{max}} = \max_{1\le i\le N}(d_i) \tag{3}$$

and the number of distances greater than 5 pixels (approximately 1% error for a resolution of 512 × 512):

$$n_{>1\%} = \frac{\Sigma_{i=1}^{N} m_i}{N} \tag{4}$$

(where $m_i = 0$ if $d_i \le 5$ and $m_i = 1$ if $d_i > 5$).

The Pratt figure of merit gives a general impression of the distances between contours; it is a relative measure and varies in the interval [0,1] where "1" means a complete match of the contours. In our case, α, which is a normalization parameter related to the size of the contours, was chosen to be 1/9 so that if all the distances d_i are equal to 3 pixels, F_{Pratt} will have a value of 0.5. The value of 3 pixels was chosen to produce a scale that allows enough discrimination among the contours drawn by the radiologists.

The mean distance also gives an integrated view of the distances between contours, while the maximum distance gives a worst case view. Finally, the number of distances greater than 1% (5 pixels in our case) provides information on the number of relevant errors and thus complements the information obtained from the previous indexes.

Another figure of merit was computed based on the similarity between the two binary images, *A* and *B*, including the areas defined by the pulmonary contours. This simple measure of similarity may be defined by:

$$\theta = \cos^{-1}\left(\frac{\mathrm{A}\cdot\mathrm{B}}{\|\mathrm{A}\|\cdot\|\mathrm{B}\|}\right) \tag{5}$$

where "." and "‖ ‖" denote the usual inner product and norm of vectors. In a Hilbert space context, θ is the angle between two binary image vectors *A* and *B*.

The correct identification of the measurement scale (24) is an important issue concerning the information provided by these figures of merit and the statistical methods that can be used. In this respect, the Pratt and θ figures of merit are measured on an ordinal scale whereas the mean error and the number of distances greater than 5 pixels are measured on a ratio scale.

Statistical Methods

As a first step in the analysis of the data obtained from the comparison among contours using all figures of merit,

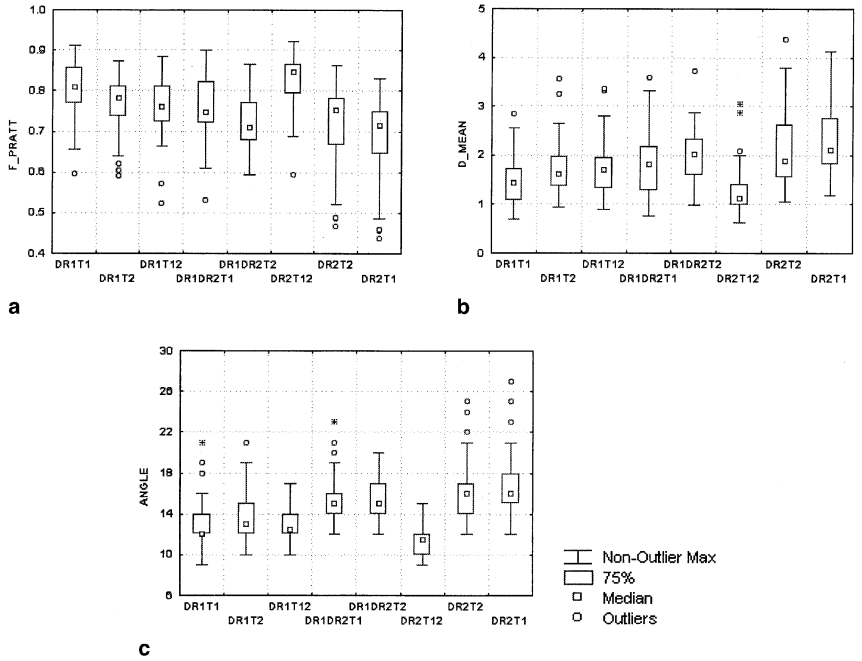

Figure 3. Box-plots for the comparison of the contours using **(a)** F_{Pratt}, **(b)** mean distance, and **(c)** angle θ, involving our algorithm and two radiologists.

we performed an exploratory data analysis (25); this analysis provided an overview of the structure of the data (showing the amplitudes, asymmetries, location, possible outliers, etc) and also some clues to the type of statistical tests to be used to test our hypothesis. The software used was *Statistica* (Statsoft, Tulsa, OK) (26).

Because the sample set did not correspond to independent experiments, nor did the data have a normal distribution, a nonparametric test was used (27). We also used multivariate data analysis (28) to assess if our algorithm is generally comparable to the six human observers as a contour detector on the used image data set.

RESULTS

Study of the Interobserver and Intraobserver Variability Involving our Algorithm and Two Expert Radiologists

The two radiologists were called DR1 and DR2, and the two contour drawing moments were called T1 and T2. Figure 3 shows the box-plots and corresponding median and quartile values for the comparison between the contours detected by our algorithm and the two radiologists in the two moments, using different figures of merit. The box-plots can be interpreted in the following way:

DRiT1–comparison between the contours detected by DRi at moment T1 to the contours detected by our algorithm, on the selected set of images;
DRiT12–comparison between the contours detected by DRi at moments T1 and T2, on the same images.

According to this notation:

DR1T12 and DR2T12 represent the intraobserver variability of experts DR1 and DR2;
DR1T1, DR1T2, DR2T1 and DR2T2 represent the interobserver variability between our algorithm and each radiologist in each moment;
DR1DR2T1 and DR1DR2T2 represent the interobserver variability between both radiologists at moments T1 and T2, respectively.

All these can be compared in Figure 3 through several figures of merit: F_{Pratt}, mean distance, and angle θ. Observing the box-plots corresponding to F_{Pratt} we note that:

1. At moment T1, DR1 is more similar to our algorithm than to DR2, because the median value of DR1T1 (median, 0.81) is higher (ie, better) than the median value of DR1DR2T1 (median, 0.75); this was con-

Table 1
Friedman ANOVA for the Comparison of the Contours using F_{Pratt}. H = 157.67, ($P < .000001$) and the Null Hypothesis is Rejected

Variable	Sum of Ranks
DR2T1	144
DR1DR2T2	186
DR2T2	203
DR1T12	272
DR1DR2T1	284
DR1T2	293
DR1T1	377
DR2T12	401

firmed using a nonparametric test for the equality of the median, the Wilcoxon test (28), which rejected the null hypothesis ($P < .00004$). Also the range of the values is smaller for DR1T1 than for DR1DR2T1. Both results suggest that the interobserver variability between the two radiologists is higher than the variability between DR1 and our algorithm;

2. At moment T2, both DR1 and DR2 are more similar to our algorithm than to each other; for instance, the median value of DR1T2 (median, 0.78) is higher than the median value of DR1DR2T2 (median, 0.70), confirmed using the Wilcoxon test ($P < .00009$).
3. DR1 is more similar to our algorithm than to himself because the median value of DR1T1 (median, 0.81) is higher (better) than the median value of DR1T12 (median, 0.76), according to the Wilcoxon test ($P < .00007$). On the other hand, the median value of DR1T2 (median, 0.78) was not considered significantly different of the median value of DR1T12, according to the Wilcoxon test ($P < .4$). The above results suggest that the interobserver variability between DR1 and our algorithm is ≤ the intraobserver variability of DR1.

These findings are not contradicted by the observation of the information obtained using the other figures of merit and were confirmed using a nonparametric method, the Friedman's two-way analysis of variance (27). The calculated H = 157.67 (with $N = 60$ and $k = 8$); under the null hypothesis (equality of medians), H has a χ^2 distribution with (k-1) degrees of freedom. In our case, for a 1% significance level (α), $\chi^2_{(7);0.01} = 18.48$; thus $H = 157.67 \gg \chi^2_{(7);0.01} = 18.48$ ($P < .000001$) and the null hypothesis is rejected. Table 1 presents the sum of ranks in ascending order.

This means that the medians are in fact significantly different, which reinforces the three observations presented above. Moreover, these observations can be confirmed through Table 1, where we can see, for instance, that the sum of ranks corresponding to DR1T1 and DR1T2 are both higher than the sum of ranks corresponding to DR1T12 (377, 299, and 272, respectively).

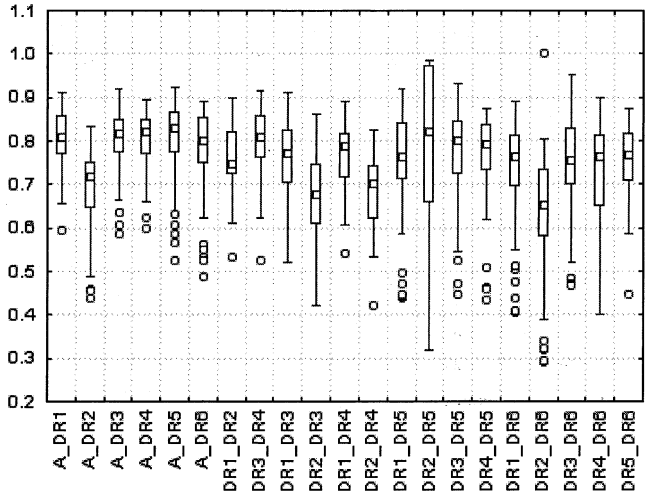

Figure 4. Box-plots comparing contours detected by the algorithm (A) and all the radiologists (DR) using F_{Pratt}.

Taking the observation of Table 1 further, we notice that all (except DR2T1) variabilities between our algorithm and each radiologist are less than (at least) the interobserver variability between the two radiologists at moment T2 (DR1DR2T2).

These and other findings that can be extracted from these results seem to indicate that, as a detector of pulmonary contours on the used set of images, our algorithm behaves as a third human observer.

Study of the Interobserver Variability Through Direct Comparison Among the Algorithm and Six Expert Radiologists

Let us generalize the previous comparison to six radiologists. DR1 . . . DR6 stand for the six radiologists and A for the algorithm. Figure 4 shows the box-plots and corresponding median and quartile values for the comparison between the contours detected by our algorithm and the six radiologists in all possible combinations using F_{Pratt} (considering that, for instance, DRi_DRj is equal to DRj_DRi, we only show one). Thus, in Figure 4, the meaning is:

A_DRi–comparison between the contours detected by our algorithm and the contours detected by DRi, on the selected set of images; it represents the interob-

server variability between our algorithm and this radiologist;

DRi_DRj—comparison between the contours detected by DRi and the contours detected by DRj, on the selected set of images; it represents the interobserver variability between these two radiologists.

Observing Figure 4 we note that the median values corresponding to situations of the type A_DRi are generally higher and more similar among them than the ones corresponding to DRi_DRj.

Performing a correspondence analysis (28) and observing the plane defined by the first two axis (which represents approximately 46% of the total inertia), we notice that our algorithm is clearly included in the main groups formed by the comparisons among all the radiologists and the algorithm. Comparisons between DR5, DR6 and DR2 seem to be isolated. This could be because DR2 had just finished his training as a radiologist and DR5 and DR6 both work in the same hospital (different from DR2).

Study of the Interobserver Variability Using a Reference Contour

We primarily show results obtained using the Pratt figure of merit because we have concluded in previous studies, and confirmed through this one, that the figures of merit (except for the maximum distance) produce consistent results, conveying the same type of information.

As a first approach, we studied the interobserver variability among all radiologists and the algorithm in a worst-case scenario. This was performed using the maximum distance figure of merit and exploratory data analysis. Figure 5 shows the box-plots of the data resulting from the comparison of the contours obtained by each detector (humans and algorithm) to the reference contours using the maximum distance. On these plots we observe a concentration of the smaller values, some outliers for all detectors (corresponding to images that should be analyzed) and median values for all detectors between 5.4 and 9.9 pixels; these values can be considered low for images of 512 × 512 pixels. Thus, even in this case all the detectors (including our algorithm) seem to have a good performance for the used image data set.

As a second approach, we studied the variability between the reference and all radiologists as well as the algorithm using the Pratt figure of merit and exploratory data analysis. In this study, we included the contours drawn by all the radiologists (DR1 to DR6) in first time, the contours drawn by DR1 and DR2 the second time (as

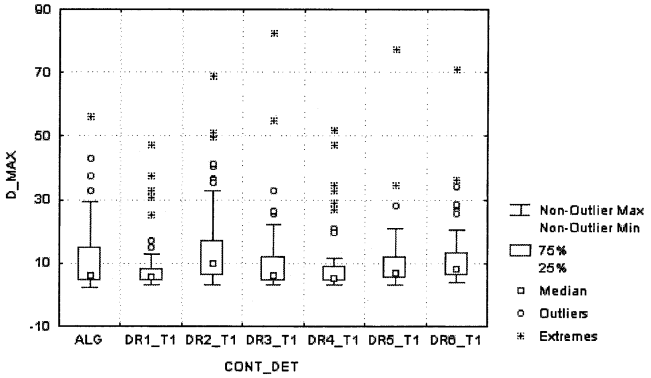

Figure 5. Box-plots corresponding to the comparison (to the reference) of the contours detected by each detector (DR1 to DR6 and the algorithm ALG) using the maximum distance figure of merit.

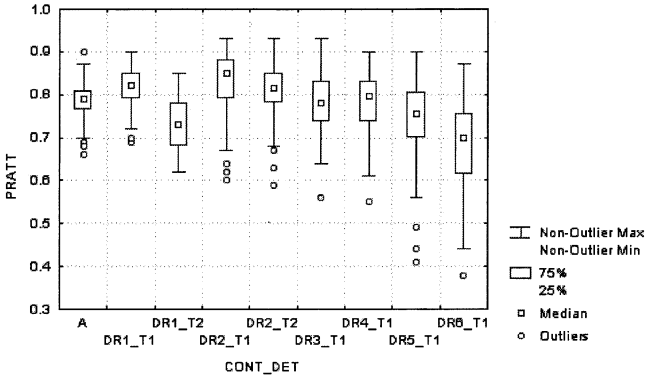

Figure 6. Box-plots corresponding to the comparison (to the reference) of the contours detected by each radiologist, in the first time (DR1_T1 to DR6_T1), two radiologists in the second time (DR1_T2 and DR2_T2) and the algorithm (A) using F_{Pratt}.

DR1_T2 and DR2_T2), as well as the contours obtained using our algorithm (A). Observing Figure 6, which shows the box corresponding to these data, we notice that the median value obtained for our algorithm is quite similar to the value for radiologist DR4_T1, higher than the values for radiologists DR1_T2, DR3_T1, DR5_T1, DR6_T1 and lower than the values for radiologists DR1_T1, DR2_T1, DR2_T2. This indicates that our algorithm produced, for the used image set, contours more similar to the reference than a significant part of the radiologists.

The above result suggested that we should further explore the relation among the performance of our algorithm as a detector to the performance of all the radiologists. Thus, we used clustering analysis (28), which closely associated our algorithm with DR1_T1 as shown by the dendogram plot of Figure 7; this means that, in this context, our algorithm is more similar to radiologist DR1

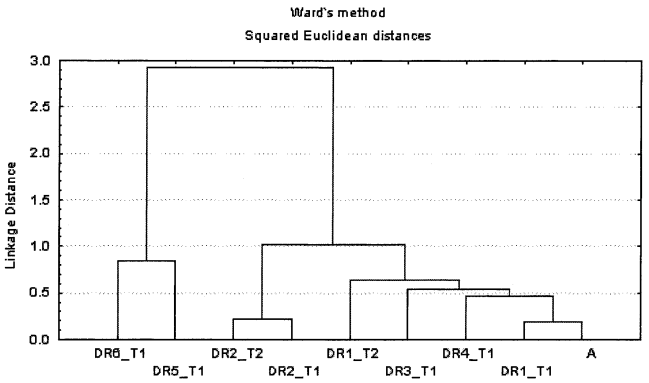

Figure 7. Dendogram plot (clustering analysis) showing all the radiologists in the first time (DR1_T1 to DR6_T1), two radiologists in the second time (DR1_T2 and DR2_T2), and the algorithm (A) using F_{Pratt}.

than he is to himself in different moments, namely DR1_T1 and DR1_T2. This conclusion was already obtained in the previous study through direct comparison among radiologists and algorithm.

A confirmation of this result was obtained through the use of another method of multivariate data analysis. Figure 8 shows the projection on the plane defined by the first two axes (approximately 66% of the total inertia) of a correspondence analysis. Observing this figure, we notice that our algorithm is clearly included in a group of four radiologists (DR1_T1, DR1_T2, DR3_ T1, DR4_T1), radiologists DR5 and DR6 form another group and DR2 is isolated between the two groups. Note that the same conclusion could be drawn from the dendogram of Figure 7. This could be related, as observed in the previous study, to the facts that radiologists DR5 and DR6 work in the same department, (different from the others) and perhaps use different segmentation criteria, radiologist DR2 has just finished his training as a radiologist and all the others have a much larger experience. To obtain a global average view of the distance between detected contours and the reference, we used the mean-distance figure of merit and angle θ, and we obtained a confirmation of the results previously found through the Pratt figure of merit (20).

CONCLUSIONS

In this article we propose a methodology to the quantitative evaluation of the performance of a pulmonary contour segmentation algorithm involving the study of interobserver and intraobserver variability.

Making accurate, unbiased estimates or comparisons of performance is, in general, a very difficult task. However, some guidelines are known to facilitate it (13,22,29). For our case, we considered the following guidelines useful:

- Report results on common test datasets;
- Use test datasets different from those used to train the segmentation method;
- Use an adequate methodology to choose the test datasets and clearly state it (eg, the inclusion and exclusion criteria and the determination of the sample size);
- Choose carefully and define clearly the observers and methods used to obtain the ground-truth;
- Let the observers operate in the same conditions;
- Clearly specify the performance metric (figures of merit) used;
- Correctly identify the measurement scales, which determine the kind of statistical methods that could be used;
- Choose hypothesis tests compatible with the quality indexes used and clearly justify it (as the chosen α and β and if the test is one- or two-tailed);
- Use nonparametric tests if the data is categorical, the statistical distribution of the data is unknown (or known and not suitable for parametric methods) or the sample size is small;
- Use paired test if possible (if all the methods can be applied to the same image).

We present results concerning the interobserver variability among six radiologists and the algorithm, using two different approaches:

Through the direct comparison of the contours detected by this algorithm to the contours hand-drawn by six radiologists;

Through a comparison to a reference contour (obtained from the hand-drawn contours) used as a surrogate ground truth.

This last approach is easier to generalize to a greater number of radiologists; however, it is necessary to further investigate what is the most correct way of computing the reference contour when radiologists use different segmentation criteria.

All the comparisons were made using several figures of merit. While the Pratt figure of merit, the mean distance, and the angle θ produced consistent results conveying the same type of information, an integrated view of the distances between contours, maximum distance is useful for worst-case scenarios.

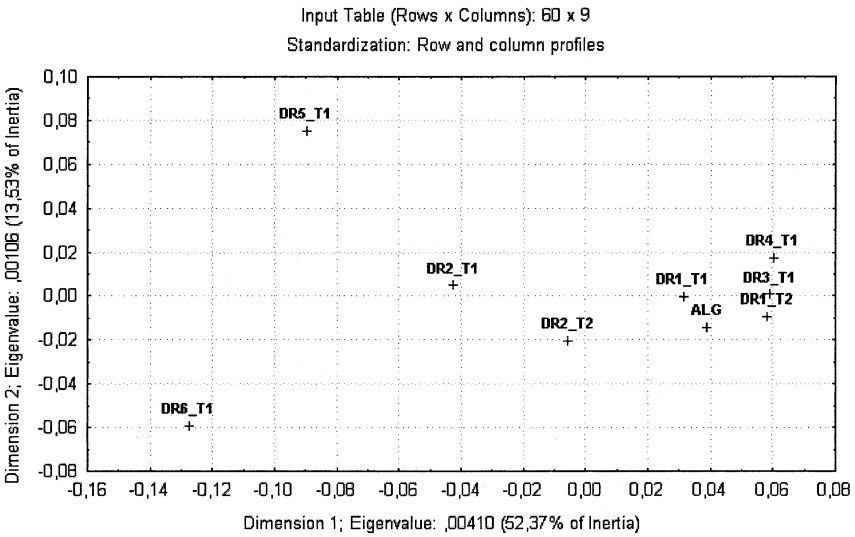

Figure 8. Correspondence analysis plot showing all the radiologists in the first time (DR1_T1 to DR6_T1), two radiologists in the second time (DR1_T2 and DR2_T2), and the algorithm (ALG) using F_{Pratt}.

We also assessed the intraobserver variability of two radiologists to have a measure of the level of interobserver variability that is expected and has to be accepted.

We believe this methodology is general enough to be applicable to many other problems of segmentation on medical images, in spite of the fact that it was developed for this specific application.

Concerning the performance of our segmentation algorithm, the results presented allow us to conclude that it is possibly as good a lung contour detector, in most thoracic CT images with diagnostic value, as any of the six radiologists. This assertion is mainly based on the fact that it exhibits a greater "agreement" to any of the radiologists than the radiologists among them, in the used image set. This is true, with a few exceptions, for images with complex vascular patterns crossing the interface between the mediastinic and pulmonary fields.

ACKNOWLEDGMENT

The authors express their gratitude to the following radiologists for drawing contours: Dr Pedro Agostinho from University Hospital of Coimbra, Dr Rui Pinho e Melo and Dr Jorge Pinho e Melo from CENTAC–Center of Computed Tomography, Aveiro; Dr Anabela Fidalgo and Dr Fernando Figueiredo from the Imagiology Department at the Hospital Infante D. Pedro, Aveiro. The authors are also grateful to an anonymous reviewer for his pertinent comments and suggestions.

REFERENCES

1. Robb WL. Perspective on the first 10 years of the CT scanner industry. Acad Radiol 2003; 10:756–760.
2. Brink J, Heiken JP, Wang G, McEnery KW, Schlueter FJ, Vannier MW. Helical CT: principles and technical considerations. Radiographics 1994; 14:887–893.
3. Li B, Christensen G, Hoffmann E, McLeannan G, Reinhardt J. Establishing a normative atlas of the human lung: intersubject warping and registration of volumetric CT images. Acad Radiol 2003; 10:255–265.
4. Brown MS, McNitt-Gray MF, Mankovich NJ, et al. Method for segmentation chest CT image data using an anatomical model: preliminary results. IEEE Trans Med Imaging 1997; 16:828–839.
5. Sonka M, Park W, Hoffman EA. Rule-based detection of intrathoracic airways trees. IEEE Trans Med Imaging 1996; 15:314–326.
6. Duryea J, Boone JM. A fully automated algorithm for the segmentation of lung fields on digital chest radiographic images. Med Phys 1995; 22:183–191.
7. Parker RP. Measurement of basic CT data. In: Moores BM, Parker RP, Pullan BR, eds. Proceedings of physical aspects of medical imaging. Manchester, UK: Wiley & Sons, 1980; 291–295.
8. Hu S, Hoffman EA, Reinhardt JM. Automatic lung segmentation for accurate quantization of volume x-ray CT images. IEEE Trans Med Imaging 2001; 20:490–498.
9. Hasegawa A, Lo S-CB, Lin J-S, Freedman MT, Mun SK. A shift-invariant neural network for the lung field segmentation in chest radiography. J VLSI Signal Process 1998; 18:241–250.
10. Silva JS, Silva A, Santos BS. Lung segmentation methods in x-ray CT images. In: Proceedings of V Ibero-American Symposium On Pattern Recognition-SIARP'2000. Lisbon, Portugal: APRP—Portuguese Association for Pattern Recognition; 2000, 583–598.
11. Silva A, Silva JS, Santos BS, Ferreira C. Fast pulmonary contour extraction in x-ray CT images: a methodology and quality assessment. In: Chen C-T, Clough AV, eds. SPIE-Medical Imaging 2001: Physiology and Function from Multidimensional Images. Bellingham, WA: SPIE, 2001;4321:216–224.
12. Chalana V, Kim Y. A methodology for evaluation of boundary detection algorithms on medical images. IEEE Trans Pattern Anal Machine Intell 1997; 16:642–652.
13. Bowyer KW. Validation of medical image analysis techniques. In: Fitzpatrick J, Sonka M, eds. Handbook of medical imaging. Vol 2. Medical

image processing and analysis (cap. X). Bellingham, WA: SPIE-The International Society for Optical Engineering, 1999; 567–606.
14. Blake A, Isard M. Active contours. London: Springer Verlag, 1998.
15. Gunn SR, Nixon MS. A robust snake implementation; a dual active contour. IEEE Trans Pattern Anal Machine Intell 1997; 19:63–68.
16. Kass M, Witkin A, Terzopoulos D. Snakes: active contour models. Int J Comput Vision 1988; 1:321–331.
17. Liang J, McInerney T, Terzopoulos D. United snakes. In: International Conference of Computer Vision—Volume 2, September 20-25, Kerkyra, Greece, 1999; 933–940.
18. Yezzi A, Kichenassamy S, Kumar A, Olver P, Tannenbaum A. A geometric snake model for segmentation of medical imagery. IEEE Trans Med Imaging 1997; 16:199–209.
19. Sivaramakrishna R, Obuchowski N, Chilcote W, Powell K. Automatic segmentation of mammographic density. Acad Radiol 2001; 8:250–256.
20. Ferreira C, Santos BS, Silva JS, Silva A. Comparison of a segmentation algorithm to six expert imagiologists in detecting pulmonary contours on x-ray CT images. In: SPIE Medical Imaging 2003: Image Perception, Observer Performance and Technology Assessment. Bellingham, WA: SPIE, 2003; 347–358.
21. Wagner R, Beiden S, Campbell G, Metz C, Sacks W. Contemporary issues for experimental design in assessment of medical imaging and computer-assist systems. In: SPIE Medical Imaging 2003: Image Perception, Observer Performance and Technology Assessment. Bellingham, WA: SPIE, 2003; 5034:213–224.
22. Altman DG. Practical statistics for medical research: London, UK: CRC Press; 1999.
23. Abdou IE, Pratt WK. Quantitative design and evaluation of enhancement/thresholding edge detectors. Proceedings IEEE 1979; 67:753–763.
24. Sachs L. Applied statistics-a handbook of techniques. New York, NY: Springer-Verlag, 1984.
25. Hoaglin D, Mosteller F, Tukey J. Understanding robust and exploratory data analysis. Wiley & Sons, 1983.
26. Statsoft. Statistica-release 5.5 for Windows. Statsoft Inc, 1999.
27. Gibbons JD. Nonparametric methods for quantitative analysis. Syracuse, NY: American Sciences Press, 1997.
28. Hair JF, Anderson RE, Tatham RL, Black WC. Multivariate data analysis with readings. Upper Saddle River, NJ: Prentice-Hall, 1995.
29. Buvat I, et al. The need to develop guidelines for the evaluation of medical image processing procedures. In: Hanson KM, ed. SPIE-Medical Imaging 1999: Image Processing. Bellingham, WA: SPIE, 1999; 1466–1477.